Computer Science Department

TECHNICAL REPORT

ON THE CONDITIONING OF POLE ASSIGNMENT

By

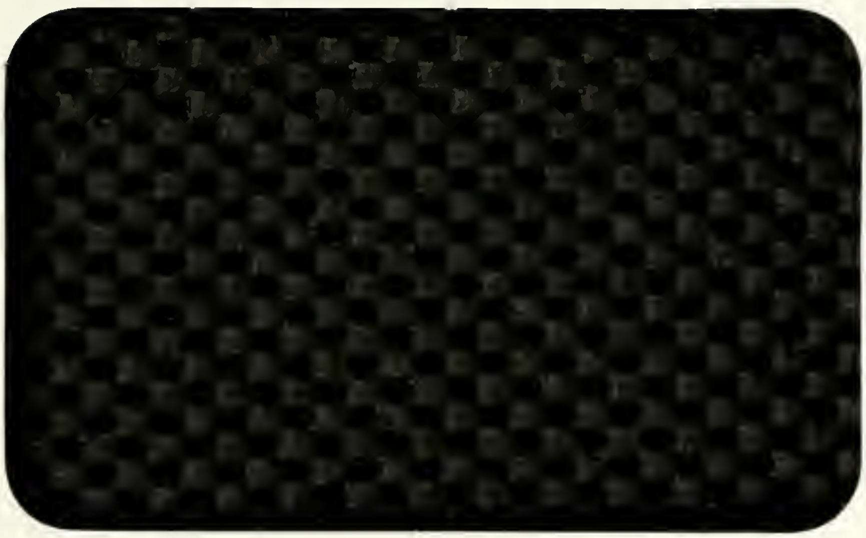
James Demmel

January 1985

Technical Report #150

NEW YORK UNIVERSITY

Department of Computer Science
Courant Institute of Mathematical Sciences
251 MERCER STREET, NEW YORK, N.Y. 10012



ON THE CONDITIONING OF POLE ASSIGNMENT

By

James Demmel

January 1985

Technical Report #150

On the Conditioning of Pole Assignment

James Demmel

Courant Institute
251 Mercer Str.
New York, NY 10012
January 15, 1985

ABSTRACT

An algorithm for the pole assignment problem has recently been given in [Kautsky,Nichols,Van Dooren]. Given the n by n matrix A, the n by m full rank matrix B ($m<n$), and the n complex numbers λ_i, the problem is to find F such that $A+BF$ has the λ_i as eigenvalues. The quality of the solution is measured by the condition number of X, the matrix of eigenvectors of $A+BF$. In this note we show that a lower bound on Cond(X) is given essentially by the reciprocal of the product of three terms: the condition number of B, the maximum condition number of any $A-\lambda_i I$, and the relative distance from the pair (A,B) to the nearest uncontrollable pair. This last term can be interpreted as follows: the lower bound on Cond(X) becomes large as the relative distance from the problem to the nearest "infinitely ill-conditioned" problem becomes small.

1. Introduction and Summary

The *pole assignment problem* is defined as follows: given real matrices A and B of dimensions n by n and n by m, respectively, B of full rank, and a set $\{\lambda_i\}$ of n complex numbers closed under complex conjugation, find a real m by n matrix F such that $A+BF$ has the λ_i as eigenvalues. It is well known that this problem has a solution for all possible self conjugate $\{\lambda_i\}$ if and only if the pair (A,B) is *controllable*, i.e. $[B \mid AB \mid A^2B \mid \cdots \mid A^{n-1}B]$ has full rank n [Wonham]. If the pair (A,B) is not controllable, then some eigenvalues of $A+BF$ will be independent of F (and be eigenvalues of A). In either case, the column space of $[B \mid AB \mid A^2B \mid \cdots \mid A^{n-1}B]$ is called the *controllable subspace* of the pair (A,B). The *robust pole assignment problem*, as defined in [Kautsky,Nichols,Van Dooren], is to find F satisfying the basic problem above

such that X, the eigenvector matrix of $A+BF$, is as well conditioned as possible. The condition number of X, $\text{Cond}(X) = \|X\| \cdot \|X^{-1}\|$, ($\| \cdot \|$ denotes the 2-norm) turns out to measure the size and sensitivity of both F and the time dependent solution of the control system defined by A and B (see [Kautsky,Nichols,Van Dooren] for details).

In this paper we consider lower bounds for $\text{Cond}(X)$. In particular we show that a lower bound on $\text{Cond}(X)$ is given essentially by the reciprocal of the product of three quantities: the condition number of B, the maximum condition number of any $A-\lambda_i I$, and the relative distance from the pair (A,B) to the nearest uncontrollable pair.

Theorem: Let $\kappa_B = \|B\|_F / \sigma_{\min}(B)$,

$$\kappa_A = \max\left\{1\ ,\ \frac{\|A\|_F}{\min_i \sigma_{\min}(A-\lambda_i)}\right\}$$

and

$$D_u = \frac{\|A_u - A\|_F}{\|A\|_F} + \frac{\|B_u - B\|_F}{\|B\|_F}\ ,$$

where (A_u,B_u) is the uncontrollable pair which minimizes D_u. Then any solution X to the robust pole assignment problem satisfies

$$\text{Cond}(X) \geq \frac{9-\sqrt{77}}{2\sqrt{n}\, D_u\, \kappa_B\, \kappa_A} > \frac{1}{9\sqrt{n}\, D_u\, \kappa_B\, \kappa_A}\ .$$

Let us interpret this result. The distance from (A,B) to the nearest uncontrollable pair is also the distance of the matrix pencil $[B \mid A-\lambda I]$ to the nearest pencil with a certain Kronecker canonical form, as we now explain. The pencil $[B \mid A-\lambda I]$ can only have two types of structure elements in its Kronecker canonical form: L_ϵ blocks

$$L_\epsilon = \begin{bmatrix} \lambda & 1 & & & \\ & \lambda & 1 & & \\ & & \cdot & \cdot & \\ & & & \lambda & 1 \end{bmatrix}$$

of dimension ϵ by $\epsilon+1$, and Jordan blocks $J_n(z_i)-\lambda I_n$ with finite eigenvalue z_i. (The presence of L_η^T blocks and infinite eigenvalues would require the coefficient of λ, $[0 \mid I]$, to have deficient row rank, which is clearly not the case.) It is known ([Wonham]) that the pair (A,B) is controllable if and only if the pencil $[B \mid A-\lambda I]$ has no regular part, i.e. no Jordan blocks. Indeed, for generic matrices A and B, the pair (A,B) will be controllable. The uncontrollable pairs (A,B) lie in a lower dimensional surface in the space of all matrix pairs. The theorem states that the scaled reciprocal of the distance (measured in the D_u norm) from a given controllable pair (A,B) to this surface will essentially determine a lower bound on Cond(X) as defined above. This idea of relating the conditioning of a problem to the distance from that problem to a lower dimensional surface of "infinitely" ill-conditioned problems arises elsewhere in numerical analysis ([Kahan],[Demmel]). The standard example of this is the condition number of a matrix with respect to inversion equaling the (scaled) reciprocal of the distance of that matrix to the nearest singular matrix (this distance equals the smallest singular value of the matrix).

The presence of κ_B and κ_A in the denominator of the lower bound mean that as B or some $A-\lambda_i$ become more ill-conditioned, the lower bound becomes smaller, and we can not guarantee an ill-conditioned solution by this technique. The solution X may still be ill-conditioned in the sense of being very sensitive to small perturbations in B and A, but this behavior does not seem to be captured by a lower bound on its condition number.

2. Proof of the Theorem

We begin by summarizing the solution of the robust pole assignment problem as described in [Kautsky,Nichols,Van Dooren]: Let x_i denote the eigenvector of $A+BF$

corresponding to λ_l. Then x_l must satisfy

$$x_l \in \mathbf{S}_l = \mathbf{N}(U_1^T(A-\lambda_l I)) \tag{1}$$

where $\mathbf{N}(\cdot)$ denotes the null space of $(\cdot)$ and

$$B = [U_0\ ,\ U_1] \cdot \begin{bmatrix} Z \\ 0 \end{bmatrix} \tag{2}$$

with $U = [U_0\ ,\ U_1]$ orthogonal and Z nonsingular. Thus, the columns of U_1 span the orthogonal complement of the column space of B. If (A,B) is controllable, then $\dim(\mathbf{S}_l)=m$ [Kautsky,Nichols,Van Dooren].

If (A,B) is not controllable, we will show that the $\mathbf{S}_l$ will generally lie in a certain $n-r$ dimensional subspace of R^n, so that the x_l will also lie in this subspace. Therefore X, whose columns are the x_l, will be singular with rank at most $n-r$. This is another way of saying that the robust pole placement problem generally has no solution for (A,B) not controllable. If (A,B) is close to being not controllable, we will show that the $\mathbf{S}_l$ will lie close to an $n-1$ dimensional subspace, so that X will be nearly singular (ill-conditioned). (If some λ_l happens already to be an eigenvalue of A, then $\dim(\mathbf{S}_l)$ will be larger than m, $\mathbf{S}_l$ will not necessarily lie in the $n-r$ dimensional subspace, and the robust pole placement problem may have a solution. This problem requires a separate analysis, which we will not do here.)

Lemma 1: If (A_u,B_u), B_u of full rank, is not controllable then there exist unitary matrices P (n by n) and

$$Q = \begin{bmatrix} I_m & 0 \\ 0 & P \end{bmatrix}$$

($n+m$ by $n+m$) such that

$$P^T[B_u \mid A_u-\lambda I]Q = \begin{bmatrix} Z_u & A_{u11}-\lambda I_m & A_{u12} & A_{u13} \\ 0 & A_{u21} & A_{u22}-\lambda I_{n-m-r} & A_{u23} \\ 0 & 0 & 0 & A_{u33}-\lambda I_r \end{bmatrix} . \tag{3}$$

Here the matrices in the first, second, third and fourth columns have m, m, $n-m-r$ and

r columns respectively, and the matrices in the first, second and third rows have m, $n-m-r$ and r rows, respectively. Furthermore, Z_u is a nonsingular matrix, and A_{u33} is upper triangular.

Proof: Choose $P = [P_1 \mid P_2 \mid P_3]$ with P_1 spanning the column space of B_u, P_2 spanning the orthogonal complement of the space spanned by P_1 within the controllable subspace of (A_u, B_u), and P_3 spanning the remainder of $\mathbf{R}^n$ (the orthogonal complement of the controllable subspace), such that A_{u33} is in Schur canonical form. The form of (3) follows as in [Wonham, Chap. 1]. Q.E.D.

From (3) we see that we can express $\mathbf{S}_l$ for the pair (A_u, B_u) as

$$\mathbf{S}_l = P \cdot \mathbf{N}(\begin{bmatrix} A_{u21} & A_{u22} - \lambda_l I_{n-m-r} & A_{u23} \\ 0 & 0 & A_{u33} - \lambda_l I_r \end{bmatrix}) .$$

Since $A_{u33} - \lambda I_r$ is the regular part of the pencil $[B_u \mid A_u - \lambda I]$, we have

Lemma 2: If λ_l is not a finite eigenvalue of the pencil $[B_u \mid A_u - \lambda I]$, then

$$\mathbf{S}_l \subseteq \operatorname{span}[P_1 \mid P_2] .$$

A sufficient but not necessary condition for λ_l not being a finite eigenvalue of the pencil $[B_u \mid A_u - \lambda I]$, is that it not be an eigenvalue of A_u.

Proof: If λ_l is not an eigenvalue of the pencil, then $A_{u33} - \lambda_l I_r$ is a nonsingular matrix, so that any vector in the null space of

$$\begin{bmatrix} A_{u21} & A_{u22} - \lambda_l I_{n-m-r} & A_{u23} \\ 0 & 0 & A_{u33} - \lambda_l I_r \end{bmatrix}$$

must have its last r components 0. Changing basis with P yields the result. Q.E.D.

Now suppose the pair (A, B) is controllable, and that the closest uncontrollable pair is (A_u, B_u). Here we measure distance using the unitarily invariant absolute norm

$$D_u = \frac{\|A_u - A\|_F}{\|A\|_F} + \frac{\|B_u - B\|_F}{\|B\|_F} , \tag{4}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. Choose P and Q as in (3) above to put the pencil $[B_u \mid A_u - \lambda I]$ into the indicated form. The next lemma describes the form of $P^T[B \mid A - \lambda I]Q$:

Lemma 3:

$$P^T[B \mid A-\lambda I]Q = \begin{bmatrix} Z_u & A_{u11}-\lambda I_m & A_{u12} & A_{u13} \\ 0 & A_{u21} & A_{u22}-\lambda I_{m-n-r} & A_{u23} \\ B_1 & A_{31} & A_{32} & A_{u33}-\lambda I_r \end{bmatrix} .$$

Furthermore, only the last rows of B_1, A_{31} and A_{32} are nonzero. Note that only the three blocks at the bottom left differ from (3).

Proof: Suppose, to the contrary, that there were differences outside the three blocks B_1, A_{31} and A_{32}. This would contradict the definition of (A_u, B_u) as being the closest uncontrollable pencil to (A, B) since by making $P^T[A_u|B_u]Q$ agree with $P^T[A|B]Q$ outside these three blocks we could find an uncontrollable pair even closer to (A, B). Similarly, if B_1, A_{31} and A_{32} had nonzero entries elsewhere than in their last rows, we could find an uncontrollable pair closer to (A, B) than (A_u, B_u) by making $P^T[A_u|B_u]Q$ agree with $P^T[A|B]Q$ outside the last rows of B_1, A_{31}, and A_{32}. Q.E.D.

Now we are ready to argue that the $\mathbf{S}_l$ for the pair (A, B) all lie close to span$[P_1|P_2]$ if the distance D_u (see (4)) from (A, B) to the nearest uncontrollable pair (A_u, B_u) is small enough, and if λ_l is far enough from the spectrum of $[B|A-\lambda I]$.

Let $R = B_1 Z_u^{-1}$, $T = (I+RR^*)^{-1/2}$, and $T_* = (I+R^*R)^{-1/2}$. Note that only the last row of R is nonzero, and that

$$\begin{aligned} \|R\|_F &\le \|B_1\|_F \, \|Z_u^{-1}\| \le \|B_1\|_F \,/\, (\sigma_{\min}(B) - \|B_1\|_F) \\ &\le D_u \, \|B\|_F \,/\, (\sigma_{\min}(B) - D_u \, \|B\|_F) = D_u \, \kappa(B) \,/\, (1 - D_u \kappa(B)) \ . \end{aligned}$$

Let

$$P_R = \begin{bmatrix} T_* & 0 & 0 \\ 0 & I_{n-m-r} & 0 \\ 0 & 0 & T \end{bmatrix} \cdot \begin{bmatrix} I_m & 0 & R^* \\ 0 & I_{m-n-r} & 0 \\ -R & 0 & I_r \end{bmatrix} .$$

It is easy to verify that P_R is unitary and that $P_R \cdot P^T[B \mid A-\lambda_l I]Q =$

$$\begin{bmatrix} T_*(Z_u+R^*B_1) & T_*(A_{u11}-\lambda_l I_m+R^*A_{31}) & T_*(A_{u12}+R^*A_{32}) & T_*(A_{u13}+R^*A_{u33}-R^*\lambda_l) \\ 0 & A_{u21} & A_{u22}-\lambda_l I_{m-n-r} & A_{u23} \\ 0 & T(A_{31}-RA_{u11})+TR\lambda_l & T(A_{32}-RA_{u12}) & T(A_{u33}-RA_{u13})-T\lambda_l \end{bmatrix}$$

so that

$$\mathbf{S}_l = P \cdot \mathbf{N}(\begin{bmatrix} A_{u21} & A_{u22}-\lambda_l I_{m-n-r} & A_{u23} \\ T(A_{31}-RA_{u11})+TR\lambda_l & T(A_{32}-RA_{u12}) & T(A_{u33}-RA_{u13})-T\lambda_l \end{bmatrix}) .$$

The unitary factor P does not change any angles between vectors so we may assume it equals I now for simplicity. Thus, $\mathbf{S}_l$ is perpendicular to the last row of

$$[A_{31}-RA_{u11}+R\lambda_l \ , A_{32}-RA_{u12} \ , A_{u33}-RA_{u13}-\lambda_l I_r] \ . \qquad (5)$$

The next lemma shows that if λ_l is not too close to an eigenvalue of $[B|A-\lambda I]$, then this row vector must lie within a small angle of $[0 \mid 0 \mid 1]$:

Lemma 4: Suppose λ_l is not close to an eigenvalue of $[B|A-\lambda I]$ in the sense that $|A_{u33_{rr}}-\lambda_l| \geq c\|A\|_F$, where $c>0$. Suppose $D_u\,\kappa(B) < 1$ and $D_u\,\kappa(B)\,/\,(1-D_u\,\kappa(B)) < c$. Then the tangent of the angle θ between the last row of (5) and $[0 \mid 0 \mid 1]$ is bounded above by

$$\tan(\theta) \leq \left(\frac{D_u\,\kappa(B)}{1-D_u\,\kappa(B)}\right) \cdot \left(\frac{6}{c-\dfrac{D_u\,\kappa(B)}{1-D_u\,\kappa(B)}}\right) .$$

Proof: $\tan(\theta)$ equals the root-sum-of-squares of the first $n-1$ components of the last row of (5) divided by the absolute value of the last component. It is easy to see that the root-sum-of-squares of the first $n-1$ components, RMS_1, is bounded above by

$$RMS_1 \leq 4\,\rho\,\|A\|_F + \rho\,|\lambda_l| \ , \qquad (7)$$

where $\rho = D_u\,\kappa(B)\,/\,(1-D_u\,\kappa(B))$, and that the absolute value of the last component,

RMS_2, is bounded below by

$$RMS_2 \geq |A_{u33_{rr}} - \lambda_l| - \rho \, \|A\|_F \ . \tag{8}$$

We consider two cases, $|\lambda_l| \leq (1+c)\|A\|_F$, and $|\lambda_l| \geq (1+c)\|A\|_F$. When $|\lambda_l| \leq (1+c)\|A\|_F$, then from (7) and (8) we have

$$\frac{RMS_1}{RMS_2} \leq \frac{\rho\,(5+c)\,\|A\|_F}{(c-\rho)\,\|A\|_F} = \frac{\rho\,(5+c)}{c-\rho} \ .$$

When $|\lambda_l| \geq (1+c)\|A\|_F$, (7) and (8) yield

$$\frac{RMS_1}{RMS_2} \leq \frac{4\,\rho\,\|A\|_F}{(c-\rho)\|A\|_F} + \frac{\rho\,|\lambda_l|}{|\lambda_l|\,(1 - \frac{1}{1+c} - \frac{\rho}{1+c})}$$

$$\leq \frac{4\,\rho}{(c-\rho)} + \frac{\rho\,(1+c)}{c-\rho} = \frac{\rho(5+\rho)}{c-\rho} \ .$$

Now bound the $5+c$ in each numerator by 6. Q.E.D.

As defined above c is difficult to estimate from the initial data A, B and λ_l. The next lemma shows how to get a simple bound on c:

Lemma 5: Let

$$\kappa_A = \max_l \left(1 \ , \ \frac{\|A\|_F}{\sigma_{\min}(A - \lambda_l I_n)}\right) \ .$$

Then

$$c \geq \frac{1}{\kappa_A} - D_u \ .$$

Proof: The 2-norm of any row of $V(A-\lambda_l I_n)V^T$ (V unitary) is at least $\sigma_{\min}(A-\lambda_l)$. In particular, the last row of $[A_{31}, A_{32}, A_{u33} - \lambda_l I_r]$ has such a norm. By the triangle inequality, the sum of the 2-norm of the last row of $[A_{31}, A_{32}]$ and $|A_{u33_{rr}} - \lambda_l|$ is at least $\sigma_{\min}(A-\lambda_l)$. Since the 2-norm of the last row of $[A_{31}, A_{32}]$ is bounded above by $D_u\|A\|_F$ we have

$$|A_{u33_{rr}} - \lambda_l| \geq \sigma_{\min}(A - \lambda_l I_n) - D_u\|A\|_F$$

or

$$\frac{|A_{u33_\pi} - \lambda_i|}{\|A\|_F} \ge \frac{\sigma_{\min}(A - \lambda_i I_n)}{\|A\|_F} - D_u$$

$$\ge \min(1\ , \frac{\sigma_{\min}(A - \lambda_i I_n)}{\|A\|_F}) - D_u$$

$$\ge \frac{1}{\kappa_A} - D_u\ .$$

Q.E.D.

We now need to translate our upper bound on $\tan(\theta)$ into a lower bound on Cond(X):

Lemma 6: Suppose the n dimensional vector $x_i \in S_i$, where the space S_i is perpendicular to the unit vector y_i. Suppose further that the angle θ_i between each y_i and a fixed unit vector e is bounded above by $\theta_{\max}$. Let X be an n by n matrix whose columns are the x_i. Then

$$\mathrm{Cond}(X) \ge \frac{1}{\sqrt{n}\ \tan(\theta_{\max})}\ .$$

Proof: Assume without loss of generality that $e=e_n$, the unit vector with a 1 in the n-th entry and zeroes elsewhere. Let $\theta(x,y)$ denote the (acute) angle between the vectors x and y. By the triangle inequality we have

$$\theta(y_i,e_n) + \theta(e_n,x_i) \ge \theta(y_i,x_i) = \pi/2$$

so that

$$\theta(e_n,x_i) \ge \pi/2 - \theta(y_i,e_n) = \pi/2 - \theta_i \ge \pi/2 - \theta_{\max}\ .$$

Writing $x_i = [x_i^T, \epsilon_i]^T$, n_i an $n-1$ vector, we see

$$\frac{|\epsilon_i|}{\|n_i\|} = \cot\,(\theta(e_n,x_i)) \le \cot(\pi/2 - \theta_{\max}) = \tan(\theta_{\max})$$

so

$$|\epsilon_i| \le \|n_i\|\ \tan(\theta_{\max}) \le \|X\|\ \tan(\theta_{\max})\ .$$

Now factor $X=LQ$, L lower triangular and Q unitary, so that L and X have the same

singular values and rows of the same norm. Thus the last row of L has the same norm as $[\epsilon_1, \ldots, \epsilon_n]$, which is in turn bounded by $\sqrt{n}\ \|X\|\ \tan(\theta_{\max})$. Since the norm of any row of L is an upper bound on its smallest singular value, we have

$$\mathrm{Cond}(X) = \mathrm{Cond}(L) \geq \frac{\|X\|}{\sqrt{n}\ \|X\|\ \tan(\theta_{\max})} = \frac{1}{\sqrt{n}\ \tan(\theta_{\max})}$$

as desired. Q.E.D.

Combining these lemmas yields the theorem.

Theorem: Let $\kappa_B = \|B\|_F \,/\, \sigma_{\min}(B)$,

$$\kappa_A = \max\left(1\ ,\ \frac{\|A\|_F}{\min\limits_i \sigma_{\min}(A-\lambda_i)}\right)$$

and

$$D_u = \frac{\|A_u - A\|_F}{\|A\|_F} + \frac{\|B_u - B\|_F}{\|B_u\|_F}\ ,$$

where (A_u, B_u) is the uncontrollable pair which minimizes D_u. Then any solution X to the robust pole assignment problem satisfies

$$\mathrm{Cond}(X) \geq \frac{9-\sqrt{77}}{2\sqrt{n}\ D_u\ \kappa_B\ \kappa_A} > \frac{1}{9n\ D_u\ \kappa_B\ \kappa_A}\ .$$

Proof: Let $\zeta = (9-\sqrt{77})/2$. We consider two cases, $D_u\ \kappa_A\ \kappa_B \geq \zeta$, and $D_u\ \kappa_A\ \kappa_B < \zeta$. In the first case,

$$\frac{9-\sqrt{77}}{2\sqrt{n}\ D_u\ \kappa_B\ \kappa_A} \leq \frac{1}{\sqrt{n}} \leq 1\ ,$$

so the theorem is trivially true. In the second case we have from Lemmas 4, 5 and 6 that

$$\mathrm{Cond}(X) \geq \frac{1}{\sqrt{n}\ \tan(\theta_{\max})} \geq \frac{(1 - D_u\kappa_B)\ (c - \dfrac{D_u\kappa_B}{1 - D_u\kappa_B})}{6\sqrt{n}\ D_u\ \kappa_B}$$

$$\geq \frac{1}{6\sqrt{n}} \frac{1 - D_u \kappa_A \kappa_B}{D_u \kappa_B} \left[\frac{1}{\kappa_A} - D_u - \frac{D_u \kappa_B}{1 - D_u \kappa_B} \right]$$

$$\geq \frac{1}{6\sqrt{n}} \frac{1 - D_u \kappa_B \kappa_A}{D_u \kappa_B \kappa_A} \left[1 - D_u \kappa_B \kappa_A - \frac{D_u \kappa_B \kappa_A}{1 - D_u \kappa_B \kappa_A} \right]$$

$$\geq \frac{1}{6\sqrt{n}} \frac{1 - \zeta}{D_u \kappa_B \kappa_A} \left[1 - \zeta - \frac{\zeta}{1 - \zeta} \right]$$

$$= \frac{1}{6\sqrt{n}} \frac{(1 - \zeta)^2 - \zeta}{D_u \kappa_B \kappa_A}$$

$$= \frac{\zeta}{\sqrt{n}\, D_u \kappa_B \kappa_A} \; ,$$

as desired. Q.E.D.

3. References

[Demmel] J. Demmel, "A Numerical Analyst's Jordan Canonical Form," Dissertation, Computer Science Dept., University of California at Berkeley, June 1983

[Kahan] W. Kahan, "Conserving Confluence Curbs Ill-Condition", Technical Report 6, Computer Science Dept., University of California, Berkeley, August 4, 1972

[Kautsky,Nichols,Van Dooren] J. Kautsky, N. Nichols, P. Van Dooren, "Robust Pole Assignment in Linear State Feedback," Numerical Analysis Report, Dept. of Mathematics, University of Reading, May 1984

[Wonham] W. M. Wonham, Linear Multivariable Control: a Geometric Approach, 2nd ed., Springer-Verlag, New York, 1979

NYU COMPSCI TR-150c.1
Demmel, James

On the conditioning
of pole assignment.

DATE DUE	BORROWER'S NAME